International Congress Series 1256 (2003) 1019–1024

Computer-aided polyp detection in CT colonography using an ensemble of support vector machines

Anna K. Jerebko[a,*], James D. Malley[b], Marek Franaszek[a], Ronald M. Summers[a]

[a] *Diagnostic Radiology Department, National Institutes of Health, Bethesda, MD 20892, USA*
[b] *Center for Information Technology, National Institutes of Health, Bethesda, MD 20892, USA*

Received 11 February 2003; received in revised form 11 February 2003; accepted 13 February 2003

Abstract

In this paper, we propose a new classification scheme for computer-aided detection (CAD) of colonic polyps in CT colonography (CTC). The scheme involves an ensemble of support vector machines (SVMs) for classification, a smoothed leave-one-out (SLOO) cross-validation method for obtaining error estimates, and the use of a bootstrap aggregation method for training and model selection. Our use of an ensemble of SVM classifiers with bagging (bootstrap aggregation), built on different feature subsets, is intended to improve classification performance when compared to single SVMs and to reduce the number of false positive detections. The bagging technique has the effect of a virtual increase in the training set size and, as a consequence, also helps to reduce the bias of error estimates when combined with a leave-one-out cross-validation approach. The bootstrap-based model selection technique is used for tuning the SVM parameters.

Keywords: Computer-aided diagnosis; Virtual colonoscopy; Colon polyp; Support vector machine; Validation

1. Introduction

CT colonography (CTC) is a less invasive and cheaper alternative to fiber optic colonoscopy. Although CTC is considerably more sensitive in polyp detection than other popular tests such as sigmoidoscopy and fecal occult blood test [1], it shows reduced

* Corresponding author. Present address: Siemens Medical Solutions USA, Inc., 51 Valley Stream Parkway, Malvern, PA 19355, USA. Tel.: +1-610-448-1828; fax: +1-610-448-1701.
E-mail address: Anna.Jerebko@siemens.com (A.K. Jerebko).

0531-5131/03 © 2003 Published by Elsevier Science B.V.
doi:10.1016/S0531-5131(03)00314-5

sensitivity compared to conventional colonoscopy. The perceptual error and fatigue due to large interpretation time are among the reasons for relatively low sensitivity rate [2]. Computer-aided diagnosis (CAD) may help reduce interpretation time in CT colonography and improve the sensitivity. Development of CAD algorithms for polyp detection in CT colonography requires accurate discrimination between true positive and false positive detections of different kinds, caused by: stool, prominent folds, noise, breathing artifacts, etc. The true polyps also have a wide variety of sizes and shapes, and this suggests using a large number of features for classification.

A wide variety of feature selection methods is available in the literature. In particular, the genetic algorithm (GA)-based method described in Ref. [3] is suitable for use with the support vector machines (SVMs) ensemble classifier suggested in this paper. In this particular study, we used expert knowledge and ROC analysis for individual features to construct our feature sets.

2. Methods

2.1. Colon segmentation and polyp candidate selection

The first step of our algorithm is colon segmentation by region growing with manual seed selection. The surface geometry information is then collected from a 3D reconstruction of the colon. From this, we declare the polyp candidate sites by applying a filter based on surface-derived geometric features such as sphericity, Gaussian and mean curvatures, number of vertices on the lesion surface, size of the lesion, tissue density and colonic wall thickness. Detection thresholds developed in earlier work were used [4]. This primary candidate selection procedure leaves approximately 16 candidate lesions per colon and still includes almost all true positive detections that could be seen on CT images, with false positives caused by retained stool, prominent folds, streak artifacts, etc.

2.2. A committee of SVMs

Many additional features are used to build the classifier for precise discrimination between real polyps and false positive detections. We add standard deviations and minimum and maximum values of the features listed above which helps to characterize texture of the tissue inside the polyp and rule out some of the false positives caused by residual stool. In addition, we calculate compactness, some volumetric characteristics of the lesion and the curvature of the ‘‘neck’’ of prospective polyp candidates.

The approach we propose in this paper is to divide the original, large set of features into several smaller subsets and then use a combination of SVM classifiers each processing a small number of input features. The subsets of variables are weighted by their effectiveness calculated on the basis of the training and test sample misclassification rates. The final, committee classification for each prospective polyp is based on the majority vote of the classifiers having effectiveness higher then a predefined threshold. The reduced feature set we obtained in this way contains only nine features, combined in four different subsets. We note that some features may be used in two or more subsets.

Each individual SVM is obtained using a first-degree polynomial kernel [5].

Consider first constructing an SVM based on the training data [5], which consists of N pairs (x_1, y_1), (x_2, y_2), …, (x_N, y_N), for p-dimensional features (predictors) $x_i \in \Re^p$ and outcomes $y_1 \in \{+1, -1\}$ (where $+1$ corresponds to polyp, -1 to non-polyp).

Define a linear decision boundary (hyperplane) by

$$\{x : f(x) = x^T\beta + \beta_0 = 0\}, \tag{1}$$

where β is a unit vector, $\|\beta\| = 1$; $f(x)$ is the signed distance from the data point x to the hyperplane.

$$G(x) = \text{sign}[f(x)] \tag{2}$$

$G(x)$ is the classification rule, such that for the test pair (x,y), the observation y is declared a polyp if $G(x)>0$, and a non-polyp if $G(x) \le 0$.

We define the *margin* for the SVM to be $2C$, where

$$C = 1/\|\beta\|. \tag{3}$$

Usually the two classes $\{y = +1\}$, $\{y = -1\}$ cannot be disjointedly separated by a hyperplane in the feature space defined by $f(x) = 0$. In this case, the support vector classifier is defined as:

$$\min\|\beta\| \text{ subject to} \begin{cases} y_i(x_i^T\beta + \beta_0) \ge 1 - \xi_i \forall i \\ \xi_i \ge 0, \sum \xi_i \le \text{constant} \end{cases} \tag{4}$$

where $\xi = (\xi_1, \xi_2, \ldots, \xi_N)$ are so-called slack variables.

Optimal estimation of $\{\beta, \beta_0\}$ is a quadratic programming problem with linear inequality constraints, and thus is a convex optimization problem, whose solution using Lagrange multipliers is described in Ref. [5].

We note that Eq. (3) is equivalent to:

$$\min \frac{1}{2}\|\beta\|^2 + \gamma \sum_{i=1}^{N} \xi_i \qquad \text{subject to } \xi_i \ge \mathbf{0}, \mathbf{y_i}(\mathbf{x_i^T}\beta x_i + \boldsymbol{\beta_0}) \ge 1 - \xi_i \forall i \tag{5}$$

where γ replaces the constant in Eq. (3).

C (see Eq. (3)) and γ are the tuning parameters in SVM classifier. Choice of C and γ is described below in Section 2.3.

2.3. Model selection and validation method

The standard tuning parameters of an SVM (C and γ) are chosen according to the user-defined sensitivity and specificity tradeoff using our model selection tool (see Fig. 1). Model selection begins with a bootstrap validation technique, introduced by Efron and Tibshirani in Ref. [6] that is applied to estimation of sensitivity and specificity, which they call the smoothed leave-one-out (SLOO) method. A grid search is performed over the range

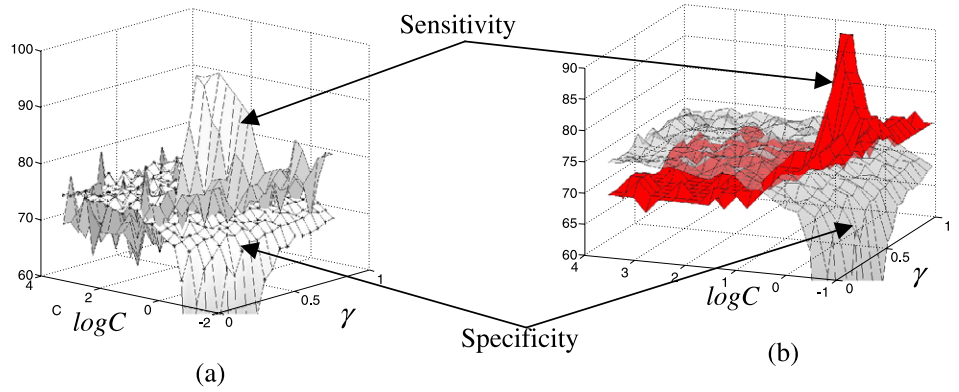

Fig. 1. SVM with 1st degree polynomial kernel. (a) Sensitivity/Specificity versus log C and γ. (b) Specificity and median-filtered sensitivity versus log C and γ.

Log$C$=[− 2,. . .,4] and γ=[0,. . .,1], and at each point of the grid, sensitivity and specificity values are estimated using the bootstrap smoothing technique: Begin with the usual leave-one-out error estimate, and set aside a single case from the N cases. Instead of using all the remaining $N-1$ cases to train and make a prediction (on the single case set aside), first do a bootstrap draw (of size N) from the $N-1$ cases, and then train all four SVMs and predict using the majority vote approach. Do boot draws many times (at least 10). Replace case, select another target case; repeat. Finally, average prediction error over all cases and boot draws.

We note that more conventional methods for estimating error rates (specificity and sensitivity) such as simple cross-validation or leave-one-out, often generate error estimates with greatly reduced variability when applied to SVM classifiers (see Ref. [5], p. 375). However, this lower observed variance generally seriously underestimates the true variability of the prediction error rates, when applied to prediction using SVMs. This is an artifact of the decision boundary constructed by any SVM, in that some (often large) percentage of the training data are not *support vectors*, and hence can be deleted from the data without affecting the placement (estimation) of the decision boundary. Hence, the leave-one-out method, for example, will typically begin by deleting a non-support vector data point, and train on a subset for which the decision boundary is exactly the same as would be found including the non-support vector data point. This leads to consistently, but artificially lower variation in the prediction error rate. A procedure such as the SLOO estimator, on the other hand, trains on multiple boot draws from the non-deleted data. On average, the bootdraws sample only a fraction of the data, specifically 0.632. In our validation approach, all the SVMs begin with a modified (subsampled) data set and each therefore generates a different decision boundary. This leads to higher, but less biased, variability of the error estimates. For this reason, we applied the SLOO estimator for error rates, rather than ordinary leave-one-out. Using the above SLOO method, we obtain two surfaces corresponding to sensitivity Se(C,γ) and specificity Sp(C,γ) (see Fig. 1). We used a median-filtering technique to further smooth the surfaces so that the optimal range of C and γ values could be obtained.

3. Experimental data and results

3.1. Data set 1

The method described above was applied to 40 CT colonography data sets obtained by prone and supine screening of 20 patients with known polyps [4]. A General Electric single detector scanner with 5-mm collimation, a 1.3-helical pitch, and a 3-mm reconstruction interval was used.

A threshold-based candidate selection algorithm was applied to 40 CT colonography data sets producing a data set suitable for further classification. The data set consisted of 47 polyps larger than 4 mm and 623 false positive detections. The model selection technique found the optimal range for Log$C$ to be $[-0.5, -0.25]$ and for γ to be [0.7, 0.8]. The ensemble sensitivity and specificity rates on test data sets (estimated with smoothed bootstrap 632+ approaches described in the previous section) were 76–78% and 70–72% (nine false positives per patient), respectively. Because the CT colonography data sets were obtained from patients with relatively poorly distended colons, we believe that the obtained misclassification rates are probably close to the best rates one could get on such data.

3.2. Data set 2

The SVM ensemble trained on data set 1 was also tested on a second data set, which contained 80 studies consisting of supine and prone screening of 40 average risk patients [7]. CT scans were done on G.E. multidetector scanners with 5-mm collimation, HQ mode, and 3-mm reconstruction interval (2 mm overlap). Ground truth was established by complete colonoscopic examination of all 40 patients. This data set consisted of CT colonography examinations obtained from patients with better colonic distension. The algorithm was able to find 13 out 15 large polyps (>1 cm) (sensitivity = 86.7%), with only three false positives per patient (1.5 per study). The sensitivity for all polyp size including medium (from 0.5 to 0.99 cm) and small polyps (less than 0.5 cm) was 75% (15 out of 20) with the same false positive rate.

In our previous experiments with this data set, the committee of four SVMs using four features each (a total of nine distinct features) allowed significant reduction in false positive rate and small improvement in sensitivity compared to average single SVM classifier built on four features (see Table 1).

We also analyzed the performance of a single large SVM with nine input features (all the features used in eight smaller SVMs composing the committee). As it can be seen in Table 1, the sensitivity, specificity, and false-positive rate of the SVM classifier built on all

Table 1
Performance of committee of SVMs compared to the average single SVM with four, and nine input features

	Average single SVM (four input features)	Single SVM (nine input features)	Committee of four SVMs
Number of false positives per study	4	4	2.6
Sensitivity (%)	71	73	81

nine features are also worse than those of committee of four SVMs. All rates in Table 1 were obtained with SLOO validation procedure.

4. Conclusion

Training an SVM ensemble on one data set and testing it on the data obtained with a different technique has shown that the classification method suggested in this paper has good generalization ability and achieves high sensitivity and a low false positive rate. The sensitivity and specificity rates are higher when the method is applied to data sets from patients having greater colonic distension. We also conclude that our model selection and improved error estimation method are effective for computer-aided polyp detection.

Acknowledgements

We thank Dr. C. Daniel Johnson of Mayo Clinic Department of Radiology for supplying CT colonography data sets.

References

[1] C.D. Johnson, W.S. Harmsen, L. Wilson, R. MacCarty, T. Welch, D. Ahlouist, Comparison of CT colonography with colonoscopy, flexible sigmoidoscopy, and fecal occult blood testing for screen-detection of colorectal polyps, RSNA 2002 scientific program, p. 444.
[2] R.M. Summers, Current concepts and future directions in computer-aided diagnosis for CT colonography, Proceedings of CARS, 16th International Congress and Exhibition, 2002, pp. 743–748.
[3] M.T. Miller, A.K. Jerebko, J.D. Malley, R.M. Summers, Optimization of the support vector classifier using genetic algorithms for variable selection, Medical Imaging SPIE 5031-12 (2003) (in press).
[4] R.M. Summers, C.D. Johnson, L.M. Pusanik, J.D. Malley, A.M. Youssef, J.E. Reed, Automated polyp detection at CT colonography: feasibility assessment in a human population, Radiology 219 (2001) 51–59.
[5] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, New York, 2001.
[6] B. Efron, R. Tibshirani, Improvements on cross-validation: the .632+ bootstrap method, Journal of the American Statistical Association 92 (438) (1997) 548–560.
[7] R.M. Summers, A. Jerebko, M. Franaszek, J.D. Malley, C.D. Johnson, Colonic polyps: complementary role of computer-aided detection in CT colonography, Radiology 225 (2002) 391–399.